Nonparametric Detection of Geometric Structures Over Networks

Shaofeng Zou, *Member, IEEE*, Yingbin Liang, *Senior Member, IEEE*, and H. Vincent Poor, *Fellow, IEEE*

***Abstract*—Nonparametric detection of the possible existence of an anomalous structure over a network is investigated. Nodes corresponding to the anomalous structure (if one exists) receive samples generated by a distribution q, which is different from a distribution p generating samples for other nodes. If an anomalous structure does not exist, all nodes receive samples generated by p. It is assumed that the distributions p and q are arbitrary and unknown. The goal is to design statistically consistent tests with probability of errors converging to zero as the network size becomes asymptotically large. Kernel-based tests are proposed based on maximum mean discrepancy, which measures the distance between mean embeddings of distributions into a reproducing kernel Hilbert space. Detection of an anomalous interval over a line network is first studied. Sufficient conditions on minimum and maximum sizes of candidate anomalous intervals are characterized in order to guarantee that the proposed test is consistent. It is also shown that certain necessary conditions must hold in order to guarantee that any test is universally consistent. Comparison of sufficient and necessary conditions yields that the proposed test is order-level optimal and nearly optimal respectively in terms of minimum and maximum sizes of candidate anomalous intervals. Generalization of the results to other networks is further developed. Numerical results are provided to demonstrate the performance of the proposed tests.**

***Index Terms*—Anomalous structure detection, consistency, maximum mean discrepancy, nonparametric test.**

I. Introduction

WE ARE interested in a type of problem, the goal of which is to detect the possible existence of an anomalous structure over a network. Each node in the network observes a random sample. An anomalous structure, if it exists, corresponds to a cluster of nodes in the network that take samples generated by a distribution q. All other nodes in the network take samples generated by a distribution p that is different from q. If there does not exist an anomalous structure, then all nodes receive samples generated by p. The distributions p and q are *arbitrary* and *unknown a priori*. Designed tests are required to distinguish between the null hypothesis (i.e., no anomalous structure exists) and the alternative hypothesis (i.e., there exists an anomalous structure). Due to the fact that the anomalous structure may be one of a number of candidate structures in the network, this is a composite hypothesis testing problem. In this paper, we first study the problem of detecting the possible existence of an anomalous interval over a line network, and then generalize our approach to higher dimensional networks.

This problem models many applications. For example, consider sensor networks in which sensors are deployed over a large area. These sensors take measurements from the environment in order to determine whether or not there is an intrusion of an anomalous object. Such intrusion typically enlarges the magnitude of the signal measured by a few sensors that are located close to the area where the intrusion occurs, whereas other sensors' measurements are still at the noise level. An alarm is then triggered if the network detects an occurrence of an intrusion based on the sensors' measurements [3]. Other applications arise in target recognition [4], detecting structural defects of bridges [5], detecting anomalous segments of DNA sequences, detecting virus infection of computer networks, and detecting anomalous spots in images [6].

As an interesting topic, detecting the possible existence of anomalous geometric structures in networks has been intensively studied in the literature as we review in Subsection I-B. However, previous studies focused on *parametric* or *semiparametric* models, which assume that samples are generated by known distributions such as Gaussian or Bernoulli distributions, or the two distributions are known to be related by a mean shift. Such parametric models may not always hold in real applications. In many cases, distributions can be arbitrary, and their difference may not be described by a simple parameter shift. The distributions may not even be known in advance. Hence, it is desirable to develop nonparametric tests that are universally applicable to arbitrary distributions.

In contrast to previous studies, we study the *nonparametric* problem of detecting an anomalous structure, in which distributions can be *arbitrary* and *unknown a priori*. In order to deal

Manuscript received November 23, 2016; revised April 27, 2017; accepted May 30, 2017. Date of publication June 22, 2017; date of current version July 24, 2017. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Wee Peng Tay. The work of S. Zou and Y. Liang was supported by the National Science Foundation CAREER Award under Grant CCF-10-26565, when S. Zou was a PhD student at Syracuse University. The work of Y. Liang was supported in part by the National Science Foundation under Grant CCF-1617789. The work of H. V. Poor was supported by the National Science Foundation under Grant CNS-1456793 and Grant CNS-1702808. This paper was presented in part at the IEEE International Workshop on Machine Learning for Signal Processing, Reims, France, September 2014, and in part at the IEEE International Conference on Acoustics, Speech and Signal Processing, Shanghai, China, March 2016. *(Corresponding author: Shaofeng Zou.)*

S. Zou is with Coordinated Science Laboratory, University of Illinois at Urbana-Champaign, Urbana, IL 61801 USA (e-mail: szou3@illinois.edu).

Y. Liang is with the Department of Electrical Engineering and Computer Science, Syracuse University, Syracuse, NY 13244 USA (e-mail: yliang06@syr.edu).

H. V. Poor is with the Department of Electrical Engineering, Princeton University, Princeton, NJ 08544 USA (e-mail: poor@princeton.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TSP.2017.2718977

with nonparametric models, we apply mean embedding of distributions into a reproducing kernel Hilbert space (RKHS) [7], [8] (see [9] for an introduction to RKHSs). The idea is to map probability distributions into an RKHS associated with an appropriate kernel such that distinguishing between two probability distributions can be carried out by evaluating the distance between the corresponding mean embeddings in the RKHS. This is valid because the mapping is shown to be injective for various kernels [10] such as Gaussian and Laplacian kernels. The main advantage of such an approach is that the mean embedding of a distribution can be easily estimated based on samples. This approach has been applied to solving the two sample problem in [11], in which the quantity of *maximum mean discrepancy (MMD)* was used as a metric of distance between mean embeddings of two distributions. In this paper, we apply MMD as a metric to construct tests for the nonparametric detection problem of interest.

We are interested in the asymptotic scenario in which the network size goes to infinity and the number of candidate anomalous structures scales with the network size. Thus, the number of sub-hypotheses under the alternative hypothesis also increases, which makes the composite hypothesis testing problem difficult. On the other hand, since the distributions are arbitrary, it is in general difficult to exploit properties of the distributions such as mean shift to detect the possible existence of an anomalous structure. Furthermore, as the network size becomes large, in contrast to parametric models in which the parameter shift can scale with the network size, the distributions here do not change with network size. Hence, it is necessary that the numbers of samples within and outside of each anomalous structure should scale with the network size fast enough in order to provide more accurate information about both distributions p and q and guarantee asymptotically small probability of error. Thus, the problem amounts to that of characterizing how the minimum and maximum sizes of all candidate anomalous structures should scale with the network size in order to consistently detect the existence of an anomalous structure.

In this paper, we adopt the following notation to express asymptotic scaling of quantities with the network size n:

- $f(n) = O(g(n))$: there exist $k, n_0 > 0$ s.t. for all $n > n_0$, $|f(n)| \leq k|g(n)|$;
- $f(n) = \Omega(g(n))$: there exist $k, n_0 > 0$ s.t. for all $n > n_0$, $f(n) \geq kg(n)$;
- $f(n) = \Theta(g(n))$: there exist $k_1, k_2, n_0 > 0$ s.t. for all $n > n_0$, $k_1 g(n) \leq f(n) \leq k_2 g(n)$;
- $f(n) = o(g(n))$: for all $k > 0$, there exists $n_0 > 0$ s.t. for all $n > n_0$, $|f(n)| \leq kg(n)$;
- $f(n) = \omega(g(n))$: for all $k > 0$, there exists $n_0 > 0$ s.t. for all $n > n_0$, $|f(n)| \geq k|g(n)|$.

A. Main Contributions

We adopt the MMD to construct nonparametric tests for various networks, which are based on kernel embeddings of distributions into an RKHS. Our main contribution lies in comprehensive analysis of the performance guarantee for the proposed tests. For the problem of detecting an anomalous interval over a line network, we show that as the network size n goes to infinity, if the minimum size $I_{\min}$ of candidate anomalous intervals satisfies $I_{\min} = \Omega(\log n)$, and the maximum size $I_{\max}$ of candidate anomalous intervals satisfies

$$n - I_{\max} = \Omega(\underbrace{\log \cdots\cdots \log}_{\text{arbitrary m number of logs}} n),$$

then the proposed test is consistent, i.e., the probability of error is asymptotically zero. We further derive necessary conditions on $I_{\min}$ and $I_{\max}$ that any test must satisfy in order to be universally consistent for arbitrary p and q. Comparison of sufficient and necessary conditions yields that the MMD-based test is order-level optimal in terms of $I_{\min}$ and nearly order-level optimal in terms of $I_{\max}$. We further generalize such analysis to other networks and obtain similar results. Our results also demonstrate the impact of geometric structures on performance guarantees for the tests.

We note that the nonparametric nature here affects the asymptotic formulation of the problem. The lower and upper bounds (such as $I_{\min}$ and $I_{\max}$ in line networks) on the sizes of all candidate anomalous structures must scale with the network size in order to guarantee enough samples in and outside the anomalous structure if it exists. This is different from parametric models in which problems can still be well posed even with a single node or the entire network being anomalous, as long as a certain distribution parameter (such as a mean shift between the two distributions) scales with the network size.

The kernel-based approach has been used to solve various machine learning problems. In recent years, Gretton and others have demonstrated the superiority of MMD as a competitive option to address the two-sample problem [11], [12], the problem of goodness of fit [13] and many other problems. In this paper, since the nature of our problem necessarily involves geometric structures in networks, the technical analysis requires an exploration of MMD for scenarios involving scaling of geometric structures and an analysis of the impact of geometry on the consistency of tests.

B. Related Work

Detecting the possible existence of an anomalous geometric structure in large networks has been extensively studied in the literature. A number of studies have focused on networks with nodes embedded in a lattice such as a one dimensional line or a square. In [14], the network is assumed to be embedded in a d-dimensional cube, and geometric structures such as line segments, disks, rectangles and ellipsoids associated with nonzero-mean Gaussian random variables need to be detected against other nodes associated with zero-mean Gaussian noise variables. A multi-scale approach was proposed and its optimality was analyzed. In [15], the detection of spatial clusters under a Bernoulli model over a two-dimensional space was studied, and a new calibration of the scan statistic was proposed, which results in optimal inference for spatial clusters. In [16], the problem of identifying a cluster of nodes with nonzero-mean values from zero-mean noise variables over a random field was studied.

Further generalization of the problem has also been studied, when network nodes are associated with a graph structure, and existence of an anomalous cluster or an anomalous subgraph of nodes needs to be detected. In [17], an unknown path corresponding to nonzero-mean variables needs to be detected out of zero-mean variables in a network with nodes connected in a graph. In [18], for various combinatorial and geometric structures of anomalous objects, conditions were established under which testing is possible or hopeless with a small risk. In [19], the cluster of anomalous nodes can either take certain geometric shapes or be connected as subgraphs. Such structures associated with nonzero-mean Gaussian variables need to be detected out of zero-mean variables. In [20] and [21], network properties of anomalous structures such as small cut size were incorporated in order to assist successful detection. More recently, in [22], the problem of detecting a connected sub-graph with elevated mean out of zero-mean Gaussian random variables was studied. An algorithm was proposed to characterize the family of all connected sub-graphs in terms of linear matrix inequalities. The minimax optimality of such an approach was further established in [23] for exponential families on one- and two-dimensional lattices.

Our problem differs from all of the above studies due to its nonparametric nature, i.e., the distributions are assumed to be unknown and arbitrary. We also note that for the problem of nonparametric detection of an anomalous interval within a line network, two nonparametric approaches based on calibration by permutation and rank-based scanning are studied in [24].

II. Preliminaries and Problem Statement

A. MMD

We provide a brief introduction to the idea of mean embedding of distributions into an RKHS [7], [8] and the metric of MMD. Suppose $\mathcal{P}$ includes a class of probability distributions, and suppose $\mathcal{H}$ is the RKHS associated with a kernel $k(\cdot,\cdot)$. We define a mapping from $\mathcal{P}$ to $\mathcal{H}$ such that each distribution $p \in \mathcal{P}$ is mapped into an element $\mu_p(\cdot)$ in $\mathcal{H}$ as follows

$$\mu_p(\cdot) = \mathbb{E}_p[k(\cdot, x)].$$

Here, $\mu_p(\cdot)$ (which maps a real variable to a real value) is referred to as the *mean embedding* of the distribution p into the Hilbert space $\mathcal{H}$. Due to the reproducing property of $\mathcal{H}$, it is clear that $\mathbb{E}_p[f] = \langle \mu_p, f\rangle_{\mathcal{H}}$, where f is any element in $\mathcal{H}$.

It is desirable that the embedding is *injective* such that each $p \in \mathcal{P}$ is mapped to a unique element $\mu_p \in \mathcal{H}$. It has been shown in [8], [10], [25], and [26] that for many RKHSs such as those associated with Gaussian and Laplacian kernels, the mean embedding is injective. In this way, many machine learning problems with unknown distributions can be solved by studying mean embeddings of probability distributions without actually estimating the distributions, e.g., [27]–[30]. In order to distinguish between two distributions p and q, [11] introduced the following quantity of maximum mean discrepancy (MMD) based on the mean embeddings μ_p and μ_q of p and q, respectively:

$$\mathrm{MMD}[p, q] := \|\mu_p - \mu_q\|_{\mathcal{H}}. \tag{1}$$

It is shown in [11] that, due to the reproducing property of the kernel,

$$\begin{aligned}\mathrm{MMD}^2[p, q] = \mathbb{E}_{x,x'}[k(x, x')] - 2\mathbb{E}_{x,y}[k(x, y)] \\ + \mathbb{E}_{y,y'}[k(y, y')],\end{aligned} \tag{2}$$

where $\mathrm{MMD}^2[p, q]$ denotes the square of $\mathrm{MMD}[p, q]$, x and x' are independent and have the same distribution p, and y and y' are independent and have the same distribution q. An unbiased estimator of $\mathrm{MMD}^2[p, q]$ based on n samples of x and m samples of y is given in [11] by

$$\begin{aligned}&\mathrm{MMD}_u^2[X, Y] \\ &= \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i}^{n} k(x_i, x_j) + \frac{1}{m(m-1)} \sum_{i=1}^{m} \sum_{j \neq i}^{m} k(y_i, y_j) \\ &\quad - \frac{2}{nm} \sum_{i=1}^{n} \sum_{j=1}^{m} k(x_i, y_j),\end{aligned} \tag{3}$$

where $X = [x_1, \ldots, x_n]$, and $Y = [y_1, \ldots, y_m]$.

B. Problem Statement

We consider a network consisting of n nodes deployed following a certain geometric structure. We use A to denote a subset of nodes that form a certain geometric structure (e.g., an interval in a line network, or a disk in a two-dimensional lattice network). Here the cardinality of A refers to the number of nodes in A, and is denoted by $|A|$. We assume that any subset A with a certain geometric structure and with cardinality between $A_{\min}$ and $A_{\max}$ can be a candidate anomalous structure, and collect all candidate anomalous structures into the following set $\mathcal{A}_n^{(a)}$:

$$\mathcal{A}_n^{(a)} = \{A : A_{\min} \le |A| \le A_{\max}\}. \tag{4}$$

As we show later, the two problem parameters $A_{\min}$ and $A_{\max}$ play an important role in determining whether the problem is well posed.

We assume that each node, say node i, is associated with a random variable, denoted by Y_i, for $1 \le i \le n$. Typically, Y_i can represent an observation received by node i. We consider the following two hypotheses. Under the *null hypothesis* H_0, Y_i for $i = 1, \ldots, n$ are independent and identically distributed (i.i.d.) random variables, and are generated from a distribution p. Under the *alternative hypothesis* H_1, there exists a geometric structure A over which Y_i (with $i \in A$) are i.i.d. generated from a distribution $q \neq p$, and otherwise, Y_i are i.i.d. generated from the distribution p. Thus, the alternative hypothesis is composite due to the fact that $\mathcal{A}_n^{(a)}$ contains multiple candidate anomalous structures. And these candidate anomalous structures differ from each other by their size or location in the network. We further assume that under both hypotheses, each node generates only one sample. Putting the problem into a context, H_0 models the scenario when the observations Y_i are background noise, and H_1 models the scenario when some Y_i (for $i \in A$) are observations activated by an anomalous event.

Again, in contrast with previous work, we assume that the distributions p and q are arbitrary and unknown *a priori*. For this problem, we are interested in the asymptotic scenario in which the number of nodes goes to infinity, i.e., $n \to \infty$. The performance of a test for such a system is captured by two types of errors. The *type I error* refers to the event that samples are generated from the null hypothesis, but the detector determines that an anomalous event occurs. We denote the probability of

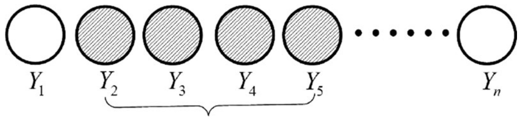


Fig. 1. A line network with an anomalous interval.

such an event as $P(H_1|H_0)$. The *type II error* refers to the case in which an anomalous event occurs but the detector claims that samples are generated from the null hypothesis. We denote the probability of such an event as $P(H_0|H_1)$. We define the following risk to measure the performance of a test for a particular pair of distributions (p, q):

$$R^{(n)} = P(H_1|H_0) + \max_{A:A\in\mathcal{A}_n^{(a)}} P(H_0|H_{1,A}). \tag{5}$$

We note that the results in this paper can also be generalized to the minimax setting in which we further take the maximum of $R^{(n)}$ over all pairs of distributions with MMD$[p,q] > \Delta$. For such a minimax setting, the threshold in our tests can be chosen to be $(1-\delta)\Delta^2$ due to the condition MMD$[p,q] > \Delta$ satisfied by (p,q), and then all sufficient conditions shown in this paper hold by replacing MMD$[p,q]$ with Δ. Furthermore, a lower bound can also be shown by choosing (p,q) satisfying MMD$[p,q] = \Delta$.

Definition 1: A test is said to be consistent if the risk $R^{(n)} \to 0$, as $n \to \infty$.

It can be seen from the definition of $\mathcal{A}_n^{(a)}$ that $A_{\min}$ and $A_{\max}$ determine the number of candidate anomalous intervals. Furthermore, if there exists an anomalous geometric structure, $A_{\min}$ determines the least number of samples generated by q and $n - A_{\max}$ determines the least number of samples generated by p. As $n \to \infty$, to guarantee asymptotically small probability of error, both $A_{\min}$ and $A_{\max}$ must scale with n to provide sufficient information about p and q in order to yield accurate distinction between the two hypotheses. This suggests that as the network becomes larger, only a large enough but not too large anomalous object can be detected. Therefore, our goal in this problem is to characterize how $A_{\min}$ and $A_{\max}$ should scale with the network size in order for a test to successfully distinguish between the two hypotheses. Such conditions on $A_{\min}$ and $A_{\max}$ can thus be interpreted as the resolution of the corresponding test.

III. LINE NETWORKS

In this section, we consider a line network consisting of nodes $1, \ldots, n$, as shown in Fig. 1. The set A in Section II-B is specialized to an *interval* I, which is a subset of consecutive indices of nodes. We consider the following set of candidate anomalous intervals:

$$\mathcal{I}_n^{(a)} = \{I : I_{\min} \le |I| \le I_{\max}\}. \tag{6}$$

A. Test and Performance

We construct a nonparametric test using the unbiased estimator in (3) and the scan statistics. For each interval I, let Y_I denote the samples in the interval I, and $Y_{\bar{I}}$ denote the samples outside the interval I. We compute $\text{MMD}^2_{u,I}(Y_I, Y_{\bar{I}})$ for all intervals $I \in \mathcal{I}_n^{(a)}$. Under the null hypothesis H_0, all samples are generated from the distribution p. Hence, for each $I \in \mathcal{I}_n^{(a)}$, $\text{MMD}^2_{u,I}(Y_I, Y_{\bar{I}})$ yields an estimate of $\text{MMD}^2[p,p]$, which is zero. Under the alternative hypothesis H_1, there exists an anomalous interval I^* in which the samples are generated from distribution q. Hence, $\text{MMD}^2_{u,I^*}(Y_{I^*}, Y_{\bar{I}^*})$ yields an estimate of $\text{MMD}^2[p,q]$, which is bounded away from zero due to the fact that $p \ne q$. Based on the above understanding, we build the following test:

$$\max_{I:I\in\mathcal{I}_n^{(a)}} \text{MMD}^2_{u,I}(Y_I, Y_{\bar{I}}) \begin{cases} \ge t, & \text{reject } H_0 \\ < t, & \text{reject } H_1 \end{cases} \tag{7}$$

where t is a threshold. It is anticipated that with a sufficiently accurate estimate of MMD and an appropriate choice of the threshold t, the test in (7) should provide the desired performance.

The computational complexity of this algorithm can be substantially improved by the fast multiscale method proposed in [14]. The basic idea of the algorithm is to use dyadic intervals (see [14] and Algorithm 1 in [1] for more details) and their extensions to approximate the candidate anomalous intervals. By this approach, the number of MMDs that need to be computed can be reduced from $O(n^2)$ to $O(n^{1+\rho})$, where $\rho > 0$ is any constant.

The following theorem characterizes the performance of the proposed test.

Theorem 1: Suppose the test in (7) is applied to the nonparametric problem described in Section III. Further assume that the kernel in the test satisfies $0 \le k(x,y) \le K$ for all (x,y), where K is a positive finite constant. Consider a pair of distributions (p,q), and assume that the threshold of the test satisfies $t < \text{MMD}^2[p,q]$. Then the type I and type II errors are upper bounded respectively as follows:

$$\begin{aligned} P(H_1|H_0) &\le \sum_{I:I_{\min}\le|I|\le I_{\max}} \exp\left(-\frac{t^2|I|(n-|I|)}{8K^2 n}\right) \\ &= \sum_{I_{\min}\le i\le I_{\max}} (n-i+1)\exp\left(-\frac{t^2 i(n-i)}{8K^2 n}\right), \end{aligned} \tag{8}$$

$$P(H_0|H_{1,I}) \le \exp\left(-\frac{(\text{MMD}^2[p,q]-t)^2|I|(n-|I|)}{8nK^2}\right), \quad \text{for } I \in \mathcal{I}_n^{(a)}. \tag{9}$$

Furthermore, the test (7) is consistent if

$$I_{\min} \ge \frac{16K^2(1+\eta)}{t^2}\log n, \tag{10}$$

$$I_{\max} \le n - \frac{16K^2(1+\eta)}{t^2} \underbrace{\log\cdots\cdots\log}_{\text{arbitrary } m \text{ number of logs}} n\ , \tag{11}$$

where η is any positive constant.

Proof: See Appendix A. ■

We note that many kernels satisfy the boundedness condition required in Theorem 1, such as the Gaussian and Laplacian kernels.

The above theorem implies that to guarantee consistency of the proposed test, the minimum length $I_{\min}$ should scale with order $I_{\min} = \Omega(\log n)$. Furthermore, $n - I_{\max}$ should scale with order

$$\Omega(\underbrace{\log \cdots\cdots \log n}_{\text{arbitrary m number of logs}})$$

which can be arbitrarily slow. Hence, the number of candidate anomalous intervals in the set $\mathcal{I}_n^{(a)}$ is $\Theta(n^2)$, which is of the same order as the total number of intervals. Hence, in the sense of order, not many intervals are excluded from being anomalous.

It can be seen that the conditions (10) and (11) on $I_{\min}$ and $I_{\max}$ are asymmetric. This can be understood by the upper bound (8) on the type I error, which is a sum over all candidate anomalous intervals with lengths between $I_{\min}$ and $I_{\max}$. Due to the specific geometric structure of the line network, as the length $|I|$ increases, the number of candidate anomalous intervals with length $|I|$ equals $n - |I| + 1$ and decreases as $|I|$ increases. Although the term $\exp(-\frac{t^2 i(n-i)}{8K^2 n})$ in (8) is symmetric over i with respect to $\frac{n}{2}$, the entire term $(n - i + 1)\exp(-\frac{t^2 i(n-i)}{8K^2 n})$ is not symmetric, which yields the asymmetric conditions on $I_{\min}$ and $I_{\max}$.

Theorem 1 requires that the threshold t in the test (7) to be less than $\text{MMD}^2[p,q]$. In practice, $\text{MMD}^2[p,q]$ may or may not be known depending on the application. If it is known, then the threshold t can be set as a constant smaller than $\text{MMD}^2[p,q]$. If it is unknown, then the threshold t needs to scale to zero as n increases without bound in order to be asymptotically smaller than $\text{MMD}^2[p,q]$. We summarize these two cases in the following two corollaries.

Corollary 1: If the value $\text{MMD}^2[p,q]$ is known *a priori*, we set the threshold $t = (1-\delta)\text{MMD}^2[p,q]$ for any $0 < \delta < 1$. The test in (7) is consistent if $I_{\min}$ and $I_{\max}$ satisfy the following conditions:

$$\begin{aligned} I_{\min} &\geq \frac{16K^2(1+\eta')}{\text{MMD}^4[p,q]} \log n \\ I_{\max} &\leq n - \frac{16K^2(1+\eta')}{\text{MMD}^4[p,q]} \underbrace{\log \cdots\cdots \log n}_{\text{arbitrary m number of logs}}, \end{aligned} \tag{12}$$

where η' is any positive constant.

Corollary 1 follows directly from Theorem 1 by setting $\eta' = \frac{1+\eta}{(1-\delta)^2} - 1$.

Corollary 2: If the value $\text{MMD}^2[p,q]$ is unknown, we set the threshold t to scale with n, such that $\lim_{n\to\infty} t_n = 0$. The test in (7) is consistent if $I_{\min}$ and $I_{\max}$ satisfy the following conditions:

$$\begin{aligned} I_{\min} &\geq \frac{16K^2(1+\eta)}{t_n^2} \log n \\ I_{\max} &\leq n - \frac{16K^2(1+\eta)}{t_n^2} \underbrace{\log \cdots\cdots \log n}_{\text{arbitrary m number of logs}}, \end{aligned} \tag{13}$$

where η is any positive constant.

Corollary 2 follows directly from Theorem 1 by noting that $t_n < \text{MMD}^2[p,q]$ for n large enough.

We note that Corollary 2 holds for any t_n that satisfies $\lim_{n\to\infty} t_n = 0$. It is clear from Corollary 2 that for the case when $\text{MMD}^2[p,q]$ is unknown, $I_{\min}$ should scale with order $\omega(\log n)$, and $n - I_{\max}$ should scale with order

$$\omega(\underbrace{\log \cdots\cdots \log n}_{\text{arbitrary m number of logs}}).$$

Hence, comparison of the above two corollaries implies that prior knowledge of $\text{MMD}^2[p,q]$ is very important for the ability to identify anomalous events. If $\text{MMD}^2[p,q]$ is known, then the network can resolve an anomalous object of size $\Omega(\log n)$. However, if such knowledge is unknown, the network can resolve only larger anomalous objects of size $\omega(\log n)$.

We note that Theorem 1 and Corollaries 1 and 2 characterize conditions to guarantee test consistency for a pair of fixed but unknown distributions p and q. Hence, the conditions (10), (11), (12) and (13) depend on the underlying distributions p and q. In fact, these conditions further yield the following condition that guarantees that the proposed test will be universally consistent for any arbitrary p and q.

Proposition 1 (Universal Consistency): Consider the nonparametric problem given in Section III. Further assume the test in (7) applies a bounded kernel with $0 \leq k(x,y) \leq K$ for any (x,y), where K is finite. If we set the threshold $t = (1-\delta)\text{MMD}^2[p,q]$ when $\text{MMD}[p,q]$ is known, and set $t_n \to 0$ as $n \to \infty$ when $\text{MMD}[p,q]$ is unknown, then the test (7) is universally consistent for any arbitrary pair of p, q, if

$$\begin{aligned} I_{\min} &= \omega(\log n) \\ I_{\max} &= n - \Omega(\underbrace{\log \cdots\cdots \log n}_{\text{arbitrary m number of logs}}). \end{aligned} \tag{14}$$

Proof: This result follows from (10), (11), (12) and (13) and the fact that $\text{MMD}[p,q]$ is a constant for any given p and q. ■

B. Necessary Conditions

In Section III, Proposition 1 suggests sufficient conditions on $I_{\min}$ and $I_{\max}$ to guarantee that the proposed nonparametric test will be universally consistent for arbitrary p and q. In the following theorem, we characterize necessary conditions on $I_{\min}$ and $I_{\max}$ that any test must satisfy in order to be universally consistent for arbitrary p and q.

Theorem 2: For the nonparametric detection problem over a line network, any test must satisfy the following conditions on $I_{\min}$ and $I_{\max}$ in order to be universally consistent for arbitrary p and q:

$$\begin{aligned} &I_{\min} = \omega(\log n) \\ \text{and} \quad &n - I_{\max} \to \infty, \text{ as } n \to \infty. \end{aligned} \tag{15}$$

Proof: See Appendix B. The idea of the proof is to first lower bound the risk by the Bayes risk of a simpler problem. Then for such a problem, we show that there exist p and q (in fact Gaussian p and q) such that even the optimal parametric test is not consistent under the conditions given in the theorem.

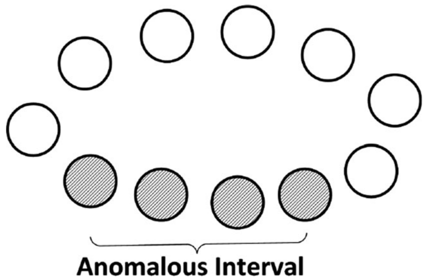


Fig. 2. A ring network with an anomalous interval.

This thus implies that under the same condition, no nonparametric test is universally consistent for arbitrary p and q. ■

Theorem 3 (Optimality): Consider the nonparametric detection problem described in Section III. If we set the threshold $t = (1-\delta)\mathrm{MMD}^2[p,q]$ when $\mathrm{MMD}[p,q]$ is known, and set $t_n \to 0$ as $n \to \infty$ when $\mathrm{MMD}[p,q]$ is unknown, the MMD-based test (7) is order-level optimal in terms of $I_{\min}$ and nearly order-level optimal in terms of $I_{\max}$ required to guarantee universal test consistency for arbitrary p and q.

Proof: It can be seen that the necessary condition on $I_{\min}$ in (15) matches the sufficient condition in (14) at the order level. Thus, the proposed test is order-level optimal in $I_{\min}$. Furthermore, the sufficient condition on $I_{\max}$ can arbitrarily slowly converge to n, which is very close to the necessary condition on $I_{\max}$. ■

IV. Generalization to Other Networks

In this section, we generalize our study to three other networks in order to demonstrate more generality of our approach. For each network, our study further demonstrates how the geometric property of the network affects the conditions required to guarantee the test consistency.

A. Detecting Intervals in Ring Networks

In this subsection, we consider a ring network (see Fig. 2), in which n nodes are located over a ring. We define an interval I to be a subset of consecutive nodes over the ring. We consider the following set of candidate anomalous intervals,

$$\mathcal{I}_n^{(a)} = \{I : I_{\min} \le |I| \le I_{\max}\}, \tag{16}$$

where $I_{\min}$ and $I_{\max}$ are minimal and maximum lengths of all candidate anomalous intervals. Despite similarities that the ring network shares with the line network, its major difference lies in that the number of candidate anomalous intervals with size k is n (which remains the same as k increases) as opposed to $n-k+1$ in the line network (which decreases as k increases). Consequently, the number of sub-hypotheses in H_1 is different. This difference is reflected in the results that we present next.

We construct the test as follows:

$$\max_{I:I\in\mathcal{I}_n^{(a)}} \mathrm{MMD}_{u,I}^2(Y_I, Y_{\bar{I}}) \begin{cases} \ge t, & \text{reject } H_0 \\ < t, & \text{reject } H_1 \end{cases} \tag{17}$$

where Y_I denotes the samples in the interval I, $Y_{\bar{I}}$ denotes the samples outside the interval I, and t is a threshold.

Theorem 4: Suppose the test (17) is applied to the problem described in Section IV-A with a bounded kernel, i.e., $0 \le k(x,y) \le K$ for all (x,y), where K is a finite positive constant. Consider a pair of distributions (p,q), and assume that the threshold of the test satisfies $t < \mathrm{MMD}^2[p,q]$. Then the type I and type II errors are bounded as follows:

$$P(H_1|H_0) \le e^{2\log n - \frac{2t^2 \min\{I_{\min}(n-I_{\min}),\, I_{\max}(n-I_{\max})\}}{16nK^2}}, \tag{18}$$

$$P(H_0|H_{1,I}) \le e^{-\frac{(\mathrm{MMD}^2[p,q]-t)^2 |I|(n-|I|)}{8nK^2}}, \text{ for } I \in \mathcal{I}_n^{(a)}. \tag{19}$$

Furthermore, the test in (17) is consistent if

$$I_{\min} \ge \frac{16K^2(1+\eta)}{t^2}\log n, \tag{20}$$

$$I_{\max} \le n - \frac{16K^2(1+\eta)}{t^2}\log n, \tag{21}$$

where η is any positive constant.

Proof: See Appendix C. ■

Comparing the above conditions with Theorem 1 suggests that although the sufficient conditions on $I_{\min}$ are the same, the conditions on $I_{\max}$ reflect an order-level difference in line and ring networks. For line networks, an anomalous interval can be close to the entire network with only a gap of length

$$\Omega(\underbrace{\log\cdots\cdots\log}_{\text{arbitrary m number of logs}} n).$$

However, for ring networks, the gap can be as large as $\Omega(\log n)$. This difference in the tests' resolution of anomalous intervals is mainly due to the difference in network geometry that further affects the error probability of the tests. By carefully comparing the two types of errors, in fact, the type II error converges to zero as the network size goes to infinity as long as the number of anomalous samples (i.e., length of anomalous intervals) and the number of typical samples (i.e., the gap between anomalous intervals and the entire network) both scale with n to infinity. Thus, the conditions for the type II error to be asymptotically zero are the same for the two types of networks. The situation is different for the type I error. The key observation is that the number of candidate anomalous intervals of size k is $n-k+1$ in a line network (which decreases as k increases), but is n in a ring network (which remains the same as k increases). This difference can be as significant as the order level if k is close to n, say $n - \Omega(\log n)$. Consequently, the type I error for a line network can be much smaller than that for a ring network, resulting in a more relaxed condition on $I_{\max}$ to guarantee consistency.

Similarly to the line network, setting the threshold t for the test (17) can be considered in two cases with and without knowledge of $\mathrm{MMD}[p,q]$. If $\mathrm{MMD}[p,q]$ is known, we set $t = (1-\delta)\mathrm{MMD}^2[p,q]$. Otherwise, t can be chosen to scale to zero as n goes to infinity. Similar results as in Corollary 1 and Corollary 2 can then be derived for the ring network.

Proposition 2 (Universal Consistency): For the nonparametric detection problem in Section IV-A, the test (17) is *universally consistent* for any arbitrary p and q, if

$$I_{\min} = \omega(\log n), \quad \text{and} \quad n - I_{\max} = \omega(\log n). \tag{22}$$

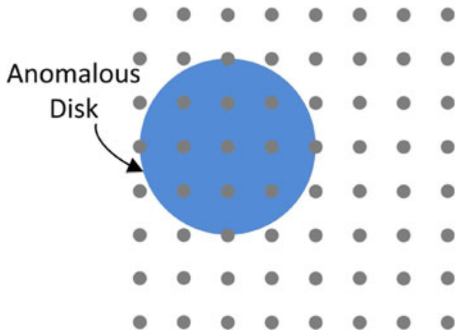


Fig. 3. Two-dimensional lattice network with an anomalous disk.

Proof: The result follows from (20) and (21) by setting $t = (1-\delta)\text{MMD}^2[p,q]$ if $\text{MMD}[p,q]$ is known, and $t_n \to 0$ as $n \to \infty$, if $\text{MMD}[p,q]$ is unknown. ■

Following arguments similar to those for the line network, it can be shown that any test must satisfy the following necessary conditions required on $I_{\min}$ and $I_{\max}$ in order to be *universally consistent* for arbitrary p and q:

$$I_{\min} = \omega(\log n), \quad \text{and} \quad n - I_{\max} \to \infty, \text{as } n \to \infty. \tag{23}$$

Theorem 5 (Optimality): Consider the problem of nonparametric detection of an interval over a ring network. If we set the threshold $t = (1-\delta)\text{MMD}^2[p,q]$ when $\text{MMD}[p,q]$ is known, and set $t_n \to 0$ as $n \to \infty$ when $\text{MMD}[p,q]$ is unknown, then the MMD-based test (17) is order-level optimal in terms of $I_{\min}$ required to guarantee universal consistency for arbitrary p and q.

Proof: Combining the above sufficient and necessary conditions, we have the optimality property for the test (17). ■

B. Detecting Disks in Two-Dimensional Lattice Networks

We consider a two-dimensional lattice network (see Fig. 3) consisting of n^2 nodes placed at the corner points of a lattice. Consider the following set of candidate anomalous disks with each disk centered at a certain node with integer radius:

$$\mathcal{D}_n^{(a)} = \{D : D_{\min} \le |D| \le D_{\max}\}, \tag{24}$$

where $|D|$ denotes the number of nodes within the disk D, $D_{\min} := \min_{D \in \mathcal{D}_n^{(a)}} |D|$ and $D_{\max} := \max_{D \in \mathcal{D}_n^{(a)}} |D|$. The goal is to detect the possible existence of an anomalous disk over the lattice network. Towards this end, we build the following test:

$$\max_{D: D \in \mathcal{D}_n^{(a)}} \text{MMD}^2_{u,D}(Y_D, Y_{\overline{D}}) \begin{cases} \ge t, & \text{reject } H_0 \\ < t, & \text{reject } H_1 \end{cases} \tag{25}$$

where Y_D contains samples within the disk D, and $Y_{\overline{D}}$ contains samples outside the disk D.

Theorem 6: Consider the problem described in Section IV-B, where the test (25) is applied with a bounded kernel, i.e., $0 \le k(x,y) \le K$, for all (x,y), where K is a finite positive constant. Then for a pair of distributions (p,q), and with $t < \text{MMD}^2[p,q]$, the type I error can be bounded as follows:

$$P(H_1|H_0) \le e^{3\log n - \frac{2t^2 \min\{D_{\min}(n^2 - D_{\min}), D_{\max}(n^2 - D_{\max})\}}{16 n^2 K^2}}, \tag{26}$$

and the type II error can be bounded as follows:

$$P(H_0|H_{1,D}) \le e^{-\frac{(\text{MMD}^2[p,q]-t)^2 |D|(n^2 - |D|)}{8n^2K^2}} \text{ for } D \in \mathcal{D}^{(a)}. \tag{27}$$

Furthermore, if $D_{\min}$ satisfies the following conditions:

$$D_{\min} \ge \frac{24K^2(1+\eta)}{t^2} \log n, \tag{28}$$

where η is any positive constant, then the test (25) is consistent.

Proof: Following steps similar to those for the line and ring networks, we can derive the bounds on the type I and type II errors shown in (26) and (27).

Then following (26) and (27), it is clear that if $D_{\min}$ and $D_{\max}$ satisfy the following conditions:

$$D_{\min} \ge \frac{24K^2(1+\eta)}{t^2} \log n, \tag{29}$$

$$D_{\max} \le n^2 - \frac{24K^2(1+\eta)}{t^2} \log n, \tag{30}$$

where η is any positive constant, then the test (25) is consistent. It is easy to verify that the largest disk within a two-dimensional lattice network has radius $\frac{n}{2}$ and area $\frac{\pi n^2}{4} \approx 0.79n^2$. Thus, such a disk contains at most cn^2 nodes where the constant $c < 1$ for large n. This implies that the bound on $D_{\max}$ in (30) is satisfied automatically when n is large. ■

Proposition 3 (Universal Consistency): For the nonparametric detection problem in Section IV-B, the test (25) is *universally consistent* for any arbitrary p and q, if

$$D_{\min} = \omega(\log n). \tag{31}$$

Proof: The result follows from (28) by setting $t = (1-\delta)\text{MMD}^2[p,q]$ if $\text{MMD}[p,q]$ is known, and setting $t_n \to 0$ as $n \to \infty$ if $\text{MMD}[p,q]$ is unknown, and the fact that $\text{MMD}[p,q]$ can be arbitrarily close to zero. ■

Following arguments similar to those for the line network, it can be shown that any test must satisfy the following necessary condition required on $D_{\min}$ in order to be *universally consistent* for arbitrary p and q:

$$D_{\min} = \omega(\log n). \tag{32}$$

Theorem 7 (Optimality): Consider the problem of nonparametric detection of a disk over a two-dimensional lattice network. If we set the threshold $t = (1-\delta)\text{MMD}^2[p,q]$ when $\text{MMD}[p,q]$ is known, and set $t_n \to 0$ as $n \to \infty$ when $\text{MMD}[p,q]$ is unknown, the MMD-based test (25) is order-level optimal in the size of disks required to guarantee universal consistency for arbitrary p and q.

Proof: Combining the above sufficient and necessary conditions, we have the optimality property for the test (25). ■

C. Detecting Rectangles in Lattice Networks

We consider an r-dimensional lattice network consisting of n^r nodes placed at the corner points of a lattice network. Consider the following set of candidate anomalous rectangles:

$$[S_n^{(a)} := \{S = [I_1 \times I_2 \times \ldots \times I_r] : S_{\min} \le |S| \le S_{\max}\},$$

where I_i for $1 \le i \le r$ denotes an interval contained in $[1, n]$ with consecutive indices, $|S|$ denotes the number of nodes in the rectangle S, $S_{\min} := \min_{S \in \mathcal{S}_n^{(a)}} |S|$, and $S_{\max} := \max_{S \in \mathcal{S}_n^{(a)}} |S|$. The goal is to detect the possible existence of an anomalous r-dimensional rectangle. Towards this end, we build the following test:

$$\max_{S: S \in \mathcal{S}_n^{(a)}} \mathrm{MMD}_{u,S}^2(Y_S, Y_{\overline{S}}) \begin{cases} \ge t, & \text{reject } H_0 \\ < t, & \text{reject } H_1 \end{cases} \quad (33)$$

where Y_S contains samples within the rectangle S, and $Y_{\overline{S}}$ contains samples outside the rectangle S.

Theorem 8: Consider the nonparametric problem described in Section IV-C, where the test (33) is applied with a bounded kernel, i.e., $0 \le k(x, y) \le K$ for all (x, y), where K is a finite positive constant. Then for a pair of (p, q), and with $t < \mathrm{MMD}^2[p, q]$, the type I error is bounded as follows:

$$P(H_1|H_0) \le e^{2r \log n - \frac{2t^2 \min\{S_{\min}(n^r - S_{\min}), S_{\max}(n^r - S_{\max})\}}{16 n^r K^2}}, \quad (34)$$

and the type II error is bounded as follows:

$$P(H_0|H_{1,S}) \le e^{-\frac{(\mathrm{MMD}^2[p,q]-t)^2 |S|(n^r - |S|)}{8 n^r K^2}}, \text{ for } S \in \mathcal{S}^{(a)}. \quad (35)$$

Furthermore, if $S_{\min}$ satisfies the following conditions:

$$S_{\min} \ge \frac{16 r K^2 (1 + \eta)}{t^2} \log n \quad (36)$$

where η is any positive constant, then the test in (33) is consistent.

Proof: Following steps similar to those developed for line and ring networks, the bounds (34) and (35) on the type I and type II errors can be derived. It is thus clear that if

$$S_{\min} \ge \frac{16 r K^2 (1 + \eta)}{t^2} \log n \quad (37)$$

$$S_{\max} \le n^r - \frac{16 r K^2 (1 + \eta)}{t^2} \log n, \quad (38)$$

then the test is consistent. We further note the important fact that as long as the largest anomalous rectangle does not span the entire lattice network, it can at most contain $n^r - n^{r-1}$ nodes, which satisfies the condition (38) for large n. ■

Proposition 4 (Universal Consistency): For the nonparametric detection problem described in Section IV-C, the test (33) is *universally consistent* for any arbitrary p and q, if

$$S_{\min} = \omega(\log n). \quad (39)$$

Proof: The result follows from (36) by setting $t = (1 - \delta)\mathrm{MMD}^2[p, q]$ if $\mathrm{MMD}[p, q]$ is known, and setting $t_n \to 0$ as $n \to \infty$ if $\mathrm{MMD}[p, q]$ is unknown, and the fact that $\mathrm{MMD}[p, q]$ can be arbitrarily close to zero. ■

Furthermore, following arguments similar to those for the line network, it can be shown that any test must satisfy the following necessary conditions required on $S_{\min}$ and $S_{\max}$ in order to be *universally consistent* for arbitrary p and q:

$$S_{\min} = \omega(\log n). \quad (40)$$

Theorem 9 (Optimality): Consider the problem of nonparametric detection of a rectangle over a lattice network. If we set the threshold $t = (1 - \delta)\mathrm{MMD}^2[p, q]$ when $\mathrm{MMD}[p, q]$ is known, and set $t_n \to 0$ as $n \to \infty$ when $\mathrm{MMD}[p, q]$ is unknown, then the MMD-based test (33) is order-level optimal to guarantee universal consistency for arbitrary p and q.

Proof: Combining the above sufficient and necessary conditions, we have the optimality property for the test (33). ■

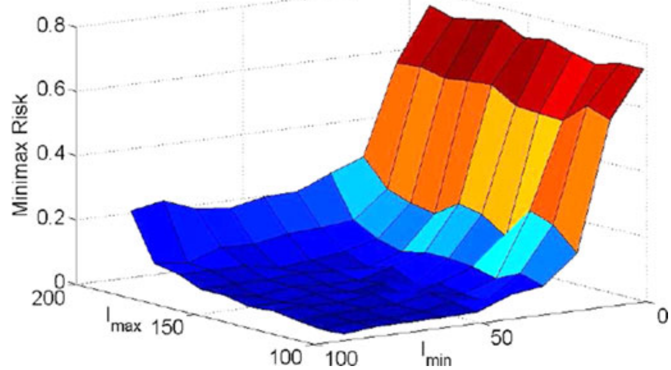


Fig. 4. Risk for a line network.

V. Numerical Results

In this section, we provide numerical results to demonstrate the performance of our tests and compare our approach with other competitive approaches.

We first simulate an intrusion detection system via a sensor network. Here, suppose sensors are deployed as a line or ring network that separates a secure area and a public area. These sensors take measurements from the environment in order to determine whether or not there is an intrusion. If there is no intrusion, the sensors receive background noise, which is modeled by the unknown typical distribution p, which is actually Gaussian with mean zero and variance one. If there is an intrusion, it activates only a few sensors within a certain area (i.e., over an interval), which take measurements from an unknown anomalous distribution q which is actually Gaussian with mean one and variance one. We set the network size $n = 200$, and use the Gaussian kernel with $\sigma = 1$. In Figs. 4 and 5, we plot the risk (normalized by 2) for line and ring networks as functions of $I_{\min}$ and $I_{\max}$. For further illustration, we also list some values of the two risk functions in Tables I and II. It can be seen from Tables I and II, and Figs. 4 and 5 that the risk functions decrease as $I_{\min}$ increases and as $I_{\max}$ decreases, which implies that our MMD based approach can detect large enough intrusion objects but not too large. This is reasonable because as $I_{\min}$ increases and as $I_{\max}$ decreases, the number of candidate anomalous intervals decreases, which reduces the difficulty of detection. The minimum numbers of samples inside and outside the anomalous interval also increase, respectively, which provides more accurate information about the distributions.

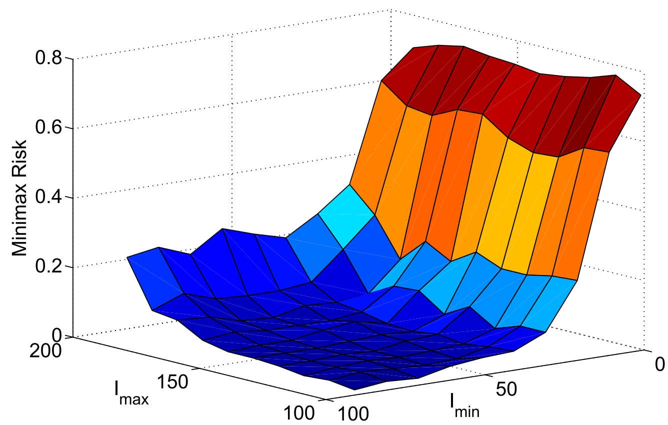


Fig. 5. Risk for a ring network.

TABLE I
RISK FOR A LINE NETWORK ($n = 200$)

$I_{\min} \setminus I_{\max}$	100	110	130	160	190
1	0.73	0.73	0.72	0.76	0.76
11	0.58	0.61	0.55	0.59	0.59
31	0.09	0.11	0.10	0.09	0.27
61	0.03	0.03	0.05	0.06	0.21
91	0.02	0.02	0.04	0.05	0.24

TABLE II
RISK FOR A RING NETWORK ($n = 200$)

$I_{\min} \setminus I_{\max}$	100	110	130	160	190
1	0.73	0.78	0.74	0.74	0.71
11	0.58	0.61	0.53	0.60	0.63
31	0.09	0.10	0.12	0.12	0.27
61	0.03	0.03	0.03	0.05	0.27
91	0.02	0.03	0.03	0.05	0.24

TABLE III
RISK FOR A LARGE LINE NETWORK ($n = 5000$)

$I_{\min} \setminus I_{\max}$	4900	4950	4980
20	0.54	0.51	0.55
50	0.02	0.02	0.54
200	0	0.01	0.52

In the second experiment, we study the performance of our test for large scale networks. We consider a line network with network size $n = 5000$. We choose the actual (but unknown) distribution p to be Gaussian with mean zero and variance one, and the actual (but unknown) anomalous distribution q to be Gaussian with mean one and variance one. We use the Gaussian kernel with $\sigma = 1$. In Table III, we list some values of the risk function of our test corresponding to various values of $I_{\min}$ and $I_{\max}$. As one can observe, the risk decreases as $I_{\min}$ increases and as $I_{\max}$ decreases. Compared to the performance for $n = 200$, although the network size becomes much larger here, the MMD test still detects small objects in the network with very small risk. This is because $I_{\min}$ and $n - I_{\max}$ only need to scale logarithmically or even slower with the network size to guarantee consistency.

In the next experiment, we compare the performance of our test with other competitive tests including Student's t-test, the Smirnov test [31], Wolf test [31], Hall test [32], kernel-based KFDA test [33] and kernel-based KDR test [34]. We consider a line network with network size $n = 100$. We set the distribution p to be Gaussian with zero mean and variance 2, and set the anomalous distribution q to be an equal mixture of Gaussian distributions $\mathcal{N}(-1, 1)$ and $\mathcal{N}(1, 1)$. Hence, the distributions p and q have the same mean and variance.

TABLE IV
COMPARISON OF NONPARAMETRIC APPROACHES OVER A LINE NETWORK VALUES OF THE RISK

$(I_{\min}, I_{\max})$	t-test	Smirnov	KFDA	KDR	MMD
(10,95)	0.90	0.92	0.70	0.66	**0.66**
(10,50)	0.88	0.90	0.51	0.56	**0.55**
(45,95)	0.89	0.93	0.54	0.43	**0.43**
(45,50)	0.83	0.62	0.06	0.06	**0.05**

TABLE V
COMPARISON OF RUNNING TIMES

	t-test	Smirnov	KFDA	KDR	MMD
Time(s)	5	41	878	265	44

In Table IV, we list some values of the risk function of our MMD-based test and other nonparametric tests corresponding to various values of $I_{\min}$ and $I_{\max}$. It can be seen that Student's t-test fails, because the test relies on a difference in mean to distinguish two distributions, which is the same in our experiment. We note that the t-test is still a powerful tool when p and q have different means. The Smirnov test estimates the cumulative distribution function (CDF) first and then takes the maximum difference of the two cumulative distribution functions as the test statistics. For continuous distributions, accurately estimating the CDF from samples requires a large amount of data, which is not feasible in our experiment. For the three kernel-based tests KFDA, KDR and MMD, the performance is very close. In particular, for large enough $I_{\min}$ and small enough $I_{\max}$, the kernel-based tests yield small risk. Among these three kernel-based tests, MMD has a slightly better performance.

We further compare the computational complexity of these algorithms via the running times they require. We set $n = 100$, $I_{\min} = 30$ and $I_{\max} = 50$. We then run these algorithms 100 times on an i7-4700 CPU, and list the average running times in Table V. It can be seen that the MMD test runs much faster than KFDA and KDR. Thus, compared to KFDA and KDR, the MMD based approach is computationally efficient. And it is much easier to implement and analyze. Furthermore, although the MMD test runs slower than the t-test and similarly to the Smirnov test, its performance (as demonstrated in Table IV) is much better than that of these two tests.

We now compare the performance of our MMD based test with the parametric approach [14]. We consider the line network. We set the network size $n = 100$, the distribution p to be Gaussian with mean zero and variance one, and the anomalous distribution q to be Gaussian with mean one and variance one. We use the Gaussian kernel with $\sigma = 1$. For the parametric test, as suggested in [14], we set the threshold $t = \sqrt{2 \log n}$. In Table VI, we list some values of the risk function of our MMD-based test and the parametric test in [14]. As we can see, the parametric test in [14] has better performance than our MMD based nonparametric test, which is due to the fact that

TABLE VI
COMPARISON WITH PARAMETRIC TEST OVER A LINE NETWORK
VALUES OF THE RISK

$(I_{\min}, I_{\max})$	MMD	Parametric Test
(10,95)	0.66	0.26
(10,50)	0.56	0.26
(45,95)	0.43	0.01
(45,50)	0.06	0.01

the parametric test has more information. And as $I_{\min}$ increases and $I_{\max}$ decreases, the gap between our MMD based test and the parametric test becomes smaller because the MMD test in such a situation has sufficient samples generated by both p and q for making decisions. Furthermore, for the parametric test, as long as $I_{\min}$ is large, the risk is small. This is due to the fact that under the parametric setting, this problem is equivalent to detecting whether there is an interval with samples generated from distribution q. Hence, it is not necessary to extract information about p from the samples, which implies that $n - I_{\max}$ can be 0.

VI. CONCLUSION

We have studied the nonparametric problem of detecting the possible existence of anomalous structures over networks, in which both the typical and the anomalous distributions can be arbitrary and unknown. We have developed nonparametric tests using the MMD to measure the distance between the mean embeddings of distributions into an RKHS. We have analyzed the performance of our tests, and characterized sufficient conditions on the minimum and maximum sizes of candidate anomalous structures to guarantee their consistency. We have further derived necessary conditions and shown that our tests are order-level optimal and nearly optimal respectively in terms of the minimum and maximum sizes of candidate structures. Thus, we have shown that the MMD-based approach can be applied to various detection problems of potential interest in practical applications.

APPENDIX A
PROOF OF THEOREM 1

We first introduce McDiarmid's inequality which is useful in bounding the probability of error in our proof.

Lemma 1 (McDiarmid's Inequality): Let $f : \mathcal{X}^m \to \mathbb{R}$ be a function such that for all $i \in \{1, \ldots, m\}$, there exist $c_i < \infty$ for which

$$\sup_{X\in\mathcal{X}^m, \tilde{x}\in\mathcal{X}} |f(x_1, \ldots, x_m) - f(x_1, \ldots x_{i-1}, \tilde{x}, x_{i+1}, \ldots, x_m)| \le c_i. \quad (41)$$

Then for any probability measure P_X over m independent random variables $X := (X_1 \ldots, X_m)$, and every $\epsilon > 0$,

$$P_X\left(f(X) - E_X(f(X)) > \epsilon\right) < \exp\left(-\frac{2\epsilon^2}{\sum_{i=1}^m c_i^2}\right), \quad (42)$$

where E_X denotes the expectation over P_X.

We now derive bounds on $P(H_1|H_0)$ and $P(H_0|H_{1,I})$ for the test (7). We first have

$$\begin{aligned}\mathrm{MMD}^2_{u,I}(Y_I, Y_{\bar{I}}) &= \frac{1}{|I|(|I|-1)}\sum_{i\in I}\sum_{\substack{j\neq i\\ j\in I}} k(y_i, y_j) \\ &+ \frac{1}{(n-|I|)(n-|I|-1)}\sum_{i\notin I}\sum_{\substack{j\neq i\\ j\notin I}} k(y_i, y_j) \\ &- \frac{2}{|I|(n-|I|)}\sum_{i\in I}\sum_{j\notin I} k(y_i, y_j). \end{aligned} \quad (43)$$

Under H_0, all samples are generated from the distribution p. Hence, $\mathbb{E}[\mathrm{MMD}^2_{u,I}(Y_I, Y_{\bar{I}})] = 0$.

In order to apply McDiarmid's inequality to bound the error probabilities $P(H_1|H_0)$ and $P(H_0|H_{1,I})$, we evaluate the following quantities. There are n variables that affects the value of $\mathrm{MMD}^2_{u,I}(Y_I, Y_{\bar{I}})$. We study the influence of these n variables on $\mathrm{MMD}^2_{u,I}(Y_I, Y_{\bar{I}})$ in the following two cases. For $i \in I$, a change in y_i affects $\mathrm{MMD}^2_{u,I}(Y_I, Y_{\bar{I}})$ through the following terms:

$$\frac{2}{|I|(|I|-1)}\sum_{\substack{j\neq i\\ j\in I}} k(y_i, y_j) - \frac{2}{|I|(n-|I|)}\sum_{j\notin I} k(y_i, y_j). \quad (44)$$

For $i \notin I$, a change in y_i affects $\mathrm{MMD}^2_{u,I}(Y_I, Y_{\bar{I}})$ through the following terms:

$$\frac{2\sum_{\substack{j\neq i\\ j\notin I}} k(y_i, y_j)}{(n-|I|)(n-|I|-1)} - \frac{2\sum_{j\in I} k(y_i, y_j)}{|I|(n-|I|)}. \quad (45)$$

Since the kernel we use is bounded, i.e., $0 \le k(x, y) \le K$ for any x, y, for $i \in I$,

$$\left|\frac{2}{|I|(|I|-1)}\sum_{\substack{j\neq i\\ j\in I}} k(y_i, y_j) - \frac{2}{|I|(n-|I|)}\sum_{j\notin I} k(y_i, y_j)\right| \le \frac{2K}{|I|},$$

and for $i \notin I$,

$$\left|\frac{2\sum_{\substack{j\neq i\\ j\notin I}} k(y_i, y_j)}{(n-|I|)(n-|I|-1)} - \frac{2\sum_{j\in I} k(y_i, y_j)}{|I|(n-|I|)}\right| \le \frac{2K}{n-|I|}.$$

Hence, we have that for $i \in I$, $c_i = \frac{4K}{|I|}$, and for $i \notin I$, $c_i = \frac{4K}{n-|I|}$, where c_i serves the role in (41).

Therefore, by applying McDiarmid's inequality, we obtain

$$P_{H_0}(\mathrm{MMD}^2_{u,I}(Y_I, Y_{\bar{I}}) > t) \le \exp\left(-\frac{2t^2|I|(n-|I|)}{16nK^2}\right). \quad (46)$$

Hence,

$$
\begin{aligned}
P(H_1|H_0) &= P_{H_0}\left(\max_{I: I\in\mathcal{I}_n^{(a)}} \mathrm{MMD}_{u,I}^2(Y_I, Y_{\bar{I}}) > t\right)\\
&\overset{(a)}{\leq} \sum_{I: I\in\mathcal{I}_n^{(a)}} P_{H_0}(\mathrm{MMD}_{u,I}^2(Y_I, Y_{\bar{I}}) > t)\\
&\leq \sum_{I: I\in\mathcal{I}_n^{(a)}} \exp\left(-\frac{2t^2|I|(n-|I|)}{16nK^2}\right)\\
&= \sum_{I: I_{\min}\leq|I|\leq I_{\max}} \exp\left(-\frac{2t^2|I|(n-|I|)}{16nK^2}\right)\\
&\overset{(b)}{=} \sum_{I: I_{\min}\leq|I|\leq n-\frac{16K^2(1+\eta)}{t^2}\log n} \exp\left(-\frac{2t^2|I|(n-|I|)}{16nK^2}\right)\\
&\quad + \sum_{I: n-\frac{16K^2(1+\eta)}{t^2}\log n+1\leq|I|\leq I_{\max}} \exp\left(-\frac{2t^2|I|(n-|I|)}{16nK^2}\right)
\end{aligned}
\tag{47}
$$

where (a) is due to Boole's inequality, η in (b) is a positive constant, and the second term in (b) is equal to zero if $n-\frac{16K^2(1+\eta)}{t^2}\log n+1\geq I_{\max}$.

It can be shown that if $I_{\min}\geq\frac{16K^2(1+\eta)}{t^2}\log n$, then the first term in (47) can be bounded as follows:

$$
\begin{aligned}
&\sum_{I: I_{\min}\leq|I|\leq n-\frac{16K^2(1+\eta)}{t^2}\log n} \exp\left(-\frac{2t^2|I|(n-|I|)}{16nK^2}\right)\\
&\overset{(a)}{\leq} n^2\exp\left(-\frac{2t^2|I|(n-|I|)}{16nK^2}\right)\Bigg|_{|I|=\frac{16K^2(1+\eta)}{t^2}\log n}\\
&= \exp\left(-2\eta\log n+o(n)\right)\to 0,\ \text{as } n\to\infty,
\end{aligned}
\tag{48}
$$

where (a) is due to the fact that there are at most n^2 candidate anomalous intervals contributing to the sum, and $|I|(n-|I|)$ is minimized by the value $|I|=\frac{16K^2(1+\eta)}{t^2}\log n$ within the range of $|I|$.

We next bound the second term in (47):

$$
\sum_{n-\frac{16K^2(1+\eta)}{t^2}\log n+1\leq|I|\leq I_{\max}} e^{-\frac{2t^2|I|(n-|I|)}{16nK^2}}
\tag{49}
$$

$$
\begin{aligned}
&= \sum_{n-\frac{16K^2(1+\eta)}{t^2}\log n+1\leq|I|\leq n-\frac{16K^2(1+\eta)}{t^2}\log\log n} e^{-\frac{2t^2|I|(n-|I|)}{16nK^2}}\\
&\quad + \sum_{n-\frac{16K^2(1+\eta)}{t^2}\log\log n+1\leq|I|\leq I_{\max}} e^{-\frac{2t^2|I|(n-|I|)}{16nK^2}}
\end{aligned}
\tag{50}
$$

$$
\begin{aligned}
&\leq \left(\frac{16K^2(1+\eta)\log n}{t^2}\right)^2 e^{-\frac{2t^2(n-\tilde{I})\tilde{I}}{16nK^2}}\\
&\quad + \sum_{n-\frac{16K^2(1+\eta)}{t^2}\log\log n+1\leq|I|\leq I_{\max}} e^{-\frac{2t^2|I|(n-|I|)}{16nK^2}},
\end{aligned}
\tag{51}
$$

where $\tilde{I}=\frac{16K^2(1+\eta)}{t^2}\log\log n$, and the first term in (51) converges to zero as n goes to infinity. The second term in (51) can be bounded as

$$
\begin{aligned}
&\sum_{n-\frac{16K^2(1+\eta)}{t^2}\log\log n+1\leq|I|\leq I_{\max}} \exp\left(-\frac{2t^2|I|(n-|I|)}{16nK^2}\right)\\
&\leq \left(\frac{16K^2(1+\eta)}{t^2}\log\log n\right)^2 \exp\left(-\frac{2t^2 I_{\max}(n-I_{\max})}{16nK^2}\right)
\end{aligned}
\tag{52}
$$

which converges to zero as $n\to\infty$ if

$$
I_{\max}\leq n-\frac{16K^2(1+\eta)}{t^2}\log\log\log n.
\tag{53}
$$

In fact, the condition (53) can be further relaxed by decomposing the second term in (51) following steps similar to (50) and (51). Such a procedure can be repeated for an arbitrary number of times, say $m-2$ times, and it can be shown that (49) converges to zero as $n\to\infty$ if

$$
I_{\max}\leq n-\frac{16K^2(1+\eta)}{t^2}\underbrace{\log\cdots\log\log}_{\text{arbitrary m number of logs}} n\ .
\tag{54}
$$

Therefore, we conclude that the type I error, i.e., $P(H_1|H_0)$, converges to zero as $n\to\infty$ if the following conditions are satisfied:

$$
\begin{aligned}
I_{\min} &\geq \frac{16K^2(1+\eta)}{t^2}\log n\\
I_{\max} &\leq n-\frac{16K^2(1+\eta)}{t^2}\underbrace{\log\cdots\log\log}_{\text{arbitrary m number of logs}} n
\end{aligned}
\tag{55}
$$

for any positive integer m.

We next continue to bound the type II error $\max_{I\in\mathcal{I}_n^{(a)}} P(H_0|H_{1,I})$ as follows:

$$
\begin{aligned}
&\max_{I\in\mathcal{I}_n^{(a)}} P(H_0|H_{1,I})\\
&= \max_{I\in\mathcal{I}_n^{(a)}} P_{H_{1,I}}\left(\max_{I'\in\mathcal{I}_n^{(a)}} \mathrm{MMD}_{u,I'}^2(Y_{I'}, Y_{\bar{I}'}) < t\right)\\
&\overset{(a)}{\leq} \max_{I\in\mathcal{I}_n^{(a)}} P_{H_{1,I}}\left(\mathrm{MMD}_{u,I}^2(Y_I, Y_{\bar{I}}) < t\right)\\
&= \max_{I\in\mathcal{I}_n^{(a)}} P_{H_{1,I}}\Big(\mathrm{MMD}^2[p,q]-\mathrm{MMD}_{u,I}^2(Y_I, Y_{\bar{I}})\\
&\qquad\qquad > \mathrm{MMD}^2[p,q]-t\Big)\\
&\overset{(b)}{\leq} \max_{I\in\mathcal{I}_n^{(a)}} \exp\left(-\frac{(\mathrm{MMD}^2[p,q]-t)^2|I|(n-|I|)}{8K^2 n}\right)
\end{aligned}
\tag{56}
$$

where (a) holds by taking $I'=I$, and (b) holds by applying McDiarmid's inequality. It can be easily checked that under the condition (55),

$$\max_{I\in\mathcal{I}_n^{(a)}} P(H_0|H_{1,I}) \le e^{-\frac{(\mathrm{MMD}^2[p,q]-t)^2|I|(n-|I|)}{8K^2n}}\Bigg|_{|I|=n-\frac{16K^2(1+\eta)}{t^2}\underbrace{\log\cdots\log\log}_{\text{arbitrary m number of logs}} n} \to 0,\ \text{as } n\to\infty. \quad (57)$$

Therefore, we conclude that the condition (55) guarantees that $R^{(n)}\to 0$ as $n\to\infty$, and thus guarantees the consistency of the test (7).

Appendix B
Proof of Theorem 2

The idea is to consider the following problem which has lower risk than our original problem, and show that there exist distributions (in fact for Gaussian p and q), under which such a risk is bounded away from zero for all tests if the necessary conditions are not satisfied.

First consider the following problem, in which all candidate anomalous intervals have the same length ℓ, and hence there are in total $n-\ell+1$ candidate anomalous intervals. Furthermore, suppose the distribution p is Gaussian with mean zero and variance one, and the distribution q is Gaussian with mean $\mu>0$ and variance one. We define the risk of a test for this problem as follows:

$$R(\ell) = P(H_1|H_0) + \max_{|I|=\ell} P(H_0|H_{1,I}), \quad (58)$$

and we denote the minimum risk as $R^*(\ell)$. We further assign a uniform prior distribution over all candidate anomalous intervals under the alternative hypothesis H_1, i.e., each candidate anomalous interval has the same probability $\frac{1}{n-\ell+1}$ of occurring. Thus the Bayes risk is given by

$$R_b = P(H_1|H_0) + \frac{1}{n-\ell+1}\sum_{|I|=\ell} P(H_0|H_{1,I}), \quad (59)$$

and we use R_b^* to denote the minimum Bayes risk over all possible tests. It is clear that

$$R^*(\ell) \ge R_b^*.$$

Following the results as in [14], [17], and [18], it can be shown that if $\ell \le \frac{1}{2\mu^2}\log n$, $R_b^*\to 1$ as n goes to infinity, which further implies that $R^*(\ell)\to 1$, as $n\to\infty$, and thus any test is no better than a random guess. Since μ can be any constant, there always exist Gaussian p and q such that no test can be consistent as long as $\ell\le c\log n$, where c is any constant.

Now consider the original problem with the risk

$$R = P(H_1|H_0) + \max_{I_{\min}\le|I|\le I_{\max}} P(H_0|H_{1,I}). \quad (60)$$

It can be shown that

$$R^* \ge R^*(\ell),\ \text{if } \ell = I_{\min}$$

where R^* denotes the minimum risk over all possible tests. Based on the above argument on $R^*(\ell)$, it is clear that if $I_{\min}\le c\log n$, there exists no test such that R^* converges to zero as n goes to infinity for arbitrary distributions p and q.

Furthermore, consider the case with only one candidate anomalous interval I with length ℓ. The risk in this case is

$$R(\ell) = P(H_1|H_0) + P(H_0|H_{1,I}) \quad (61)$$

where $|I|=\ell$. It is also clear that $R^*\ge R^*(\ell)$ where $\ell = I_{\max}$. For such a simple case, the problem reduces to the two-sample problem, detecting whether the samples in the interval I and the samples outside of the interval I are generated from the same distribution. In order to guarantee $R^*(k)\to 0$ as $n\to\infty$, ℓ and $n-\ell$ should both scale with n to infinity. Thus, in order to guarantee $R^*\to 0$, as $n\to\infty$, $n-I_{\max}\to\infty$ is necessary for any test to be universally consistent. This concludes the proof.

Appendix C
Proof of Theorem 4

Following steps similar to those in Appendix A, we derive the following bound:

$$\begin{aligned}
P(H_1|H_0) &\le \sum_{I\in\mathcal{I}_n^{(a)}} \exp\left(-\frac{2t^2|I|(n-|I|)}{16nK^2}\right)\\
&\overset{(a)}{=} \sum_{i=I_{\min}}^{I_{\max}} n\exp\left(-\frac{2t^2 i(n-i)}{16nK^2}\right)\\
&\overset{(b)}{\le} \sum_{i=I_{\min}}^{I_{\max}} n\exp\left(-\frac{2t^2\min\{I_{\min}(n-I_{\min}), I_{\max}(n-I_{\max})\}}{16nK^2}\right)\\
&\overset{(c)}{\le} n^2\exp\left(-\frac{2t^2\min\{I_{\min}(n-I_{\min}), I_{\max}(n-I_{\max})\}}{16nK^2}\right)\\
&= \exp\left(2\log n - \frac{2t^2\min\{I_{\min}(n-I_{\min}), I_{\max}(n-I_{\max})\}}{16nK^2}\right),
\end{aligned}$$

where (a) is due to the fact in the ring network there are n candidate anomalous intervals with size i, (b) is due to the fact that $i(n-i)$ is lower bounded by $\min\{I_{\min}(n-I_{\min}), I_{\max}(n-I_{\max})\}$, and (c) is due to the fact that $I_{\max}-I_{\min}\le n$.

It can be checked that $P(H_1|H_0)\to 0$ as $n\to\infty$ if

$$\frac{16K^2(1+\eta)}{t^2}\log n \le I_{\min} \le I_{\max} \le n - \frac{16K^2(1+\eta)}{t^2}\log n. \quad (62)$$

Furthermore, following steps similar to those in Appendix A, we can derive an upper bound on the type II error and show that it converges to zero as $n\to\infty$ if

$$I_{\min}\to\infty, \quad n-I_{\max}\to\infty. \quad (63)$$

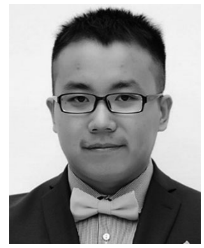
Combining the two conditions completes the proof.

References

[1] S. Zou, Y. Liang, and H. V. Poor, "A kernel-based nonparametric test for anomaly detection over line networks," in *Proc. IEEE Int. Workshop Mach. Learn. Signal Process.*, Reims, France, 2014, pp. 1–6.

[2] S. Zou, Y. Liang, and H. V. Poor, "Nonparametric detection of an anomalous disk over a two-dimensional lattice network," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, Shanghai, China, 2016, pp. 2394–2398.

[3] R. A. Kemmerer and G. Vigna, "Intrusion detection: A brief history and overview," *Computer*, vol. 35, no. 4, pp. supl27–supl30, 2002.
[4] D. Casasent and A. Ye, "Detection filters and algorithm fusion for ATR," *IEEE Trans. Image Process.*, vol. 6, no. 1, pp. 114–125, Jan. 1997.
[5] W. Liu, S. Chen, and E. Hauser, "LiDAR-based bridge structure defect detection," *Exp. Techn.*, vol. 35, no. 6, pp. 27–34, 2011.
[6] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," *ACM Comput. Surv.*, vol. 41, no. 3, p. 15, 2009.
[7] A. Berlinet and C. Thomas-Agnan, *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. Berlin, Germany: Springer Science & Business Media, 2004.
[8] B. Sriperumbudur, A. Gretton, K. Fukumizu, G. Lanckriet, and B. Schölkopf, "Hilbert space embeddings and metrics on probability measures," *J. Mach. Learn. Res.*, vol. 11, pp. 1517–1561, 2010.
[9] B. Schölkopf and A. Smola, *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. Cambridge, MA, USA: MIT Press, 2001.
[10] K. Fukumizu, A. Gretton, X. Sun, and B. Schölkopf, "Kernel measures of conditional dependence," in *Proc. Adv. Neural Inf. Process. Syst.*, 2007, pp. 489–496.
[11] A. Gretton, K. Borgwardt, M. Rasch, B. Schölkopf, and A. Smola, "A kernel two-sample test," *J. Mach. Learn. Res.*, vol. 13, pp. 723–773, 2012.
[12] K. P. Chwialkowski, A. Ramdas, D. Sejdinovic, and A. Gretton, "Fast two-sample testing with analytic representations of probability measures," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 1981–1989.
[13] K. P. Chwialkowski, H. Strathmann, and A. Gretton, "A kernel test of goodness of fit," in *Proc. Int. Conf. Mach. Learn.*, 2016, pp. 2606–2615.
[14] E. Arias-Castro, D. L. Donoho, and X. Huo, "Near-optimal detection of geometric objects by fast multiscale methods," *IEEE Trans. Inf. Theory*, vol. 51, no. 7, pp. 2402–2425, Jul. 2005.
[15] G. Walther, "Optimal and fast detection of spatial clusters with scan statistics," *Ann. Statist.*, vol. 38, no. 2, pp. 1010–1033, 2010.
[16] P. M. Pacifico, C. Genovese, I. Verdinelli, and L. Wasserman, "False discovery control for random fields," *J. Amer. Statist. Assoc.*, vol. 99, pp. 1002–1014, 2004.
[17] E. Arias-Castro, E. J. Candes, H. Helgason, and O. Zeitouni, "Searching for a trail of evidence in a maze," *Ann. Statist.*, vol. 36, no. 4, pp. 1726–1757, 2008.
[18] L. Addario-Berry, N. Broutin, L. Devroye, and G. Lugosi, "On combinatorial testing problems," *Ann. Statist.*, vol. 38, no. 5, pp. 3063–3092, 2010.
[19] E. Arias-Castro, E. J. Candes, and A. Durand, "Detection of an anomalous cluster in a network," *Ann. Statist.*, vol. 39, no. 1, pp. 278–304, 2011.
[20] J. Sharpnack, A. Rinaldo, and A. Singh, "Changepoint detection over graphs with the spectral scan statistic," in *Proc. Int. Conf. Artif. Intell. Statist.*, vol. 13, Scottsdale, AZ, 2013, pp. 545–553.
[21] J. Sharpnack, A. Rinaldo, and A. Singh, "Detecting activations over graphs using spanning tree wavelet bases," in *Proc. Artif. Intell. Statist.*, Scottsdale, AZ, 2013, pp. 536–544.
[22] J. Qian, V. Saligrama, and Y. Chen, "Connected sub-graph detection," in *Proc. Int. Conf. Artif. Intell. Statist.*, 2014, pp. 796–804.
[23] J. Qian and V. Saligrama, "Efficient minimax signal detection on graphs," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 2708–2716.
[24] E. Arias-Castro, R. M. Castro, E. Tánczos, and M. Wang, "Distribution-free detection of structured anomalies: Permutation and rank-based scans," *J. Amer. Statist. Assoc.*, to be published.
[25] B. Sriperumbudur, A. Gretton, K. Fukumizu, G. Lanckriet, and B. Schölkopf, "Injective Hilbert space embeddings of probability measures," in *Proc. Annu. Conf. Learn. Theory*, 2008, pp. 111–122.
[26] K. Fukumizu, B. Sriperumbudur, A. Gretton, and B. Schölkopf, "Characteristic kernels on groups and semigroups," in *Proc. Adv. Neural Inf. Process. Syst.*, 2009, 473–480.
[27] L. Song, A. Gretton, and K. Fukumizu, "Kernel embeddings of conditional distributions," *IEEE Signal Process. Mag.*, vol. 30, no. 4, pp. 98–111, 2013.
[28] L. Song, A. Parikh, and E. P. Xing, "Kernel embeddings of latent tree graphical models," in *Proc. Adv. Neural Inf. Process. Syst.*, Dec. 2011, pp. 2708–2716.
[29] L. Song, A. Gretton, D. Bickson, Y. Low, and C. Guestrin, "Kernel belief propagation," in *Proc. Int. Conf. Artif. Intell. Statist.*, 2011, 707–715.
[30] A. J. Smola, A. Gretton, L. Song, and B. Schölkopf, "A Hilbert space embedding for distributions," in *Algorithmic Learning Theory (Lecture Notes Comput. Sci.*, vol. 4754), E. Takimoto, Ed. Berlin, Germany: Springer, 2007.
[31] J. H. Friedman and L. C. Rafsky, "Multivariate generalizations of the Wald-Wolfowitz and Smirnov two-sample tests," *Ann. Statist.*, vol. 7, no. 4, pp. pp. 697–717, 1979.
[32] P. Hall and N. Tajvidi, "Permutation tests for equality of distributions in high-dimensional settings," *Biometrika*, vol. 89, no. 2, pp. 359–374, 2002.
[33] Z. Harchaoui, F. Bach, and E. Moulines, "Testing for homogeneity with kernel Fisher discriminant analysis," in *Proc. Adv. Neural Inf. Process. Syst.*, 2008, pp. 609–616.
[34] M. Sugiyama, T. Suzuki, and T. Kanamori, *Density Ratio Estimation in Machine Learning*. New York, NY, USA: Cambridge University Press, 2012.



Shaofeng Zou (S'14–M'16) received the B.E.(Hons.) degree from Shanghai Jiao Tong University, Shanghai, China, and the Ph.D. degree in electrical and computer engineering from Syracuse University, Syracuse, NY, USA, in 2011 and 2016, respectively. Since July 2016, he has been a Postdoctoral Research Associate at the University of Illinois at Urbana-Champaign, Urbana, IL, USA. His research interests include information theory, machine learning, and statistical signal processing.

Yingbin Liang (S'01–M'05–SM'16) received the Ph.D. degree in electrical engineering from the University of Illinois at Urbana-Champaign, Urbana, IL, USA, in 2005. From 2005 to 2007, she was working as a Postdoctoral Research Associate at Princeton University. From 2008 to 2009, she was an Assistant Professor at the University of Hawaii. Since December 2009, she has been working as an Associate Professor at Syracuse University, Syracuse, NY, USA. Her research interests include machine learning, statistical signal processing, optimization, information theory, and wireless communication and networks.

She was a Vodafone Fellow at the University of Illinois at Urbana-Champaign during 2003–2005, and received the Vodafone-U.S. Foundation Fellows Initiative Research Merit Award in 2005. In 2005, She also received the M. E. Van Valkenburg Graduate Research Award from the ECE department, University of Illinois at Urbana-Champaign. In 2009, she received the National Science Foundation CAREER Award, and the State of Hawaii Governor Innovation Award. In 2014, she received EURASIP Best Paper Award for the *EURASIP Journal on Wireless Communications and Networking*. From 2013 to 2015, she served as an Associate Editor for the Shannon Theory of the IEEE TRANSACTIONS ON INFORMATION THEORY.

H. Vincent Poor (S'72–M'77–SM'82–F'87) received the Ph.D. degree in EECS from Princeton University, Princeton, NJ, USA in 1977. From 1977 to 1990, he was on the faculty of the University of Illinois at Urbana-Champaign. Since 1990, he has been on the faculty at Princeton, where he is the Michael Henry Strater University Professor of Electrical Engineering. From 2006 to 2016, he served as the Dean of Princeton's School of Engineering and Applied Science. His research interests include the areas of information theory and signal processing, and their applications in wireless networks and related fields. Among his publications in these areas is the recent book *Information Theoretic Security and Privacy of Information Systems* (Cambridge, U.K.: Cambridge University Press, 2017).

Dr. Poor is a Member of the National Academy of Engineering and the National Academy of Sciences, and a Foreign Member of the Royal Society. In 1990, he served as the President of the IEEE Information Theory Society, and from 2004 to 2007 as the Editor-in-Chief of the IEEE TRANSACTIONS ON INFORMATION THEORY. He received the Technical Achievement and Society Awards of the IEEE Signal Processing Society in 2007 and 2011. Recent recognition of his work includes the 2017 IEEE Alexander Graham Bell Medal and a D.Sc. *honoris causa* from Syracuse University awarded in 2017. He was elected as a Foreign Member of the National Academy of Engineering of Korea and an Honorary Member of the National Academy of Sciences of Korea, both in 2017.